

Abstract

In epistemic community, people are said to be selected on their knowledge contribution to the project (articles, codes, etc.) However, the socialization process is an important factor for inclusion, sustainability as a contributor, and promotion. Finally, what does matter to be promoted? being a good contributor? being a good animator? knowing the boss? We explore this question looking at the process of election for administrator in the English Wikipedia community. We modeled the candidates according to their revisions and/or social attributes. These attributes are used to construct a predictive model of promotion success, based on the candidates's past behavior, computed thanks to a random forest algorithm.

Our model combining knowledge contribution variables and social networking variables successfully explain 78% of the results which is better than the former models. It also helps to refine the criterion for election. If the number of knowledge contributions is the most important element, social interactions come close second to explain the election. But being connected with the future peers (the admins) can make the difference between success and failure, making this epistemic community a very social community too.

Social Interactions vs Revisions, What is important for Promotion in Wikipedia?

Romain Picot-Clément^{1,2}, Cécile Bothorel^{1,2}, Nicolas Jullien^{1,3*}

January 8, 2015

1 Introduction

Mobilizing hundreds (Linux) to thousands of contributors (Wikipedia), volunteer online open projects aiming at creating new knowledge, online “communities of creation”, as named by [Rullani and Haefliger \(2013\)](#), are viewed as central in the generation of new, innovative knowledge by and for firms. But the path to successful community building is still risky and uncertain, and as for business building, most of the attempts fail, no matter how many hundreds of thousands of dollars were put into them ([Worthen, 2008](#)).

One of the key elements to develop a successful and sustainable community, as explained a quarter of century ago by Eric von Hippel ([1986](#)), is to attract enough highly competent and “committed/committing” contributors, being they named “lead users”, “core”, or “big” contributors ([Mahr and Lievens, 2012](#); [Fang and Neufeld, 2009](#)), i.e. the most productive people, who are also those with more responsibility in the management of the project ([Rullani and Haefliger, 2013](#)). How, on what criterion these core contributors are recruited to become managers is still a matter of research, something we want to address here.

If we agree with the theory of epistemic community ([Cohendet et al., 2001](#); [Edwards, 2001](#)), which stresses that those communities are project-oriented communities of experts, evaluated on their contribution in terms of knowledge, the main criterion for promotion in the different steps of the career will be their knowledge production. For Wikipedia, when projects have rules for running for administrator, they are about knowing the rules, but also about the number of edits of articles (more than 3,000 and more of one year of activity for the French Wikipedia, https://fr.wikipedia.org/wiki/Wikipédia:Candidature_au_statut_d'administrateur). [O'Mahony and Ferraro \(2007, part II\)](#), studying Open Source projects, showed that “developers who were making greater technical contributions and who were more engaged in organization building were more likely to become members of the leadership team”. (p. 1096). [Fleming and Waguespack \(2007\)](#) found the same result in their study of the Internet Engineering Task Force community.

However, [Pentzold \(2011\)](#), for Wikipedia, [von Krogh et al. \(2012\)](#), for open source, defended the idea that becoming a big contributor may be an additional

^{*}¹Institut Mines Telecom Bretagne, ²UMR CNRS 3192 Lab-STICC, ³ICI-M@rsouin, {Romain.PicotClemente, Cecile.Bothorel, Nicolas.Jullien}@telecom-bretagne.eu

step from being a regular contributor, an additional commitment, which would occur for reasons developed during the attendance of the project as the development of this sense of "community", i.e. to understand and accept the rules of the organization (Butler et al., 2008; Cardon, 2012). If we follow their argument, the social interactions with peers may be an additional requirement for being promoted, as explained by (Rullani and Haeffliger, 2013).

This article discusses whether the knowledge contribution or the social connexion matter more for being promoted in an online epistemic community, looking at the electing process of the administrators (admin) in the English Wikipedia, called the Request for Adminship (RfA). Any user participating on Wikipedia can request to become administrator, but every candidate is not elected, as RfA has about a 4 in 10 chance of being successful.

The article is organized as follows: in section 2 a review of the literature used to construct our framework of investigation, in section 3 the formulation of our hypotheses, in section 4 the data collection strategy (choice of the community and definition of the questions), in section 5 the results. We discuss the consequences of this work, its limits and future research in section 6 before concluding.

2 Literature review

The fact that these projects are made possible by the aggregation of various motivations and levels of involvement is classic for social theories, being the critical mass theory, which regards the construction of collective action (Oliver et al., 1985; Marwell and Oliver, 1993), or the theory of (knowledge) commons (Ostrom, 1990; Hess and Ostrom, 2006).

More precisely, Shah (2006) showed, in the case of open source, that long-term participants enjoyed programming and interacting with the rest of the community (i.e., labeled as "hobbyists"), whereas short-term participants were typically driven by an immediate need for software (i.e., use value). Theoretical analyses of incentives, in software projects (Foray and Zimmermann, 2001; Lerner and Tirole, 2002) or in wikis (Forte and Bruckman (2005), using Latour and Woolgar's (1979) analysis of science "cycles of credit"), estimate that the other main vector for participation, beside being able to have feedbacks and improvements on the pieces of knowledge proposed, is the quest for reputation. Applied works on Wikipedia (Nov, 2007; Yang and Lai, 2010; Zhang and Zhu, 2011), professional electronic networks (Wasko and Faraj, 2005; Jullien et al., 2011), and open source software (Shah, 2006; Scacchi, 2007), confirm that peer recognition, whether it be professional or community recognition, is a main motive for contribution amongst the main contributors, in addition to intrinsic factors (personal enjoyment and satisfaction from helping by sharing their knowledge).

In the same time, the role of a person evolves over time in the communities of creation (Von Krogh et al., 2003; Jensen and Scacchi, 2007), as do the reasons why s/he participates (Von Krogh et al., 2003). Regarding Wikipedia, for instance, it has been showed that there is process of specialization of the editors (Iba et al., 2010; Welser et al., 2011), with multipurpose wikipedians, able to participate in the redaction of an article, and wikipedians focused on the global editing, i.e. the coördination and the organization of the project.

Consequently, contributors may develop their social connections, interacting with peers and this may drive social recognition. Beyond a certain level of contribution, one can even wonder if these social interactions matter more than additional knowledge production for receiving peer recognition. More precisely, we wonder if it is possible to predict the promotion of a user according to her activity, separating this activity into knowledge production (i.e. “edits” of the articles, or “revisions”, for Wikipedia, “commits” for open source software), and social activity. This is what we tested here, using the English Wikipedia as case study.

3 Choice of the case study

Although Open Source initiatives are numerous, in various industries ([Balka et al., 2009](#)), the main open knowledge project outside the computer industry is to be found in the encyclopedia editing project known as Wikipedia. It has become one of the most successful knowledge production projects ever, with more than 4 million articles for the English version and more than one million visits per day, and is seen as a model for knowledge management theory ([McAfee, 2006](#); [Hasan and Pfaff, 2006](#)).

In addition Wikipedia has a process of election for some of the managing task, the administrator position, where social connections and knowledge production skills seem to matter, and provides a complete set of data regarding this process.

The Administrators have more rights than normal users on Wikipedia; they can (un)block specific users from editing pages, they can do some special actions on pages like (un)protecting from editing, (un)deleting, renaming, reporting vandalism, etc. In Wikipedia, every registered user can request to be promoted administrator. Nevertheless, not all RfA are successful, it is not easy to become an administrator. Precise rules for applying vary from one language to another. As already said, we will focus on the English Wikipedia and its rules here, where the RfA process is as follows. First, candidate creates a page dedicated to the request. During seven days which can be assimilated to a campaign, the candidate is questioned and his characteristics are studied by anybody from the community. Every user can vote for or against the candidate, and can change her mind at any moment during this period. At the end of the campaign, following the votes and the discussions on the RfA page, special users called “bureaucrats” gives their verdict about accepting or rejecting the RfA. There is no objective threshold on the percentage of support votes needed to be elected by bureaucrats. Nevertheless, it appears that a candidate is more likely to pass if he achieves at least 80% support. With less than 70% support, the candidate is generally not promoted. From 2001 to 2008, on the English Wikipedia, 2794 users have requested to become administrators. Among them, 1248 requests have been accepted, so that the success rate is about 44.7%.

Previous studies shed light on the entanglement of the knowledge production skills and of the social skills for being promoted. Considering the knowledge skills, [Leskovec et al. \(2010\)](#) showed that voters are more likely to give positive votes when candidates are more active than them in terms of “edits”. the flip side of the coin is that they also showed that voters who have spoken to the candidate before the RfA tend to cast a positive ballot. Still regarding the importance of socialization, [Lee et al. \(2012\)](#) considering the implicit social network from talk

pages of users showed that voter’s vote was highly correlated with the one of his neighbors in the network: voters were more likely to participate to elections involving their contacts; influent users participating in a vote can influence the final result. And the probability of a positive vote, with an accuracy of 84%, is function of the intensity of the voter’s relations (i.e. co-editions, discussions, co-revert¹) with the candidate (Jankowski-Lorek et al., 2013).

Closer to what we want to study here, Burke and Kraut (2008) proposed a model to predict RfA results according to the criteria put forward by the Guide to RfAs², where are described the criteria RfA evaluators look for in nominees, mainly based on accounting candidate’s activity: Strong edit history, with Edit summaries (explaining what they did when editing), and High quality of articles, Varied experience, User interaction, Helping with chores (i. e. already working on admin tasks such as discussing articles for delation), Trustworthiness, Observing consensus, and having various experiences in terms of editing, user interaction, etc. Their model’s accuracy reached 75.6%. However, as the authors mentioned, their measure of some variables, the less obvious, such as trustworthiness, if simple to compute, are quite naïve. On the other hand, as it has been proved in the case of measuring articles’ quality, it is not sure that increasing the complexity of the measure improves its accuracy very much³.

Moreover, regarding the question we address, here, they did not measure the respective influence of the edits and of the social interaction on RfA result. They did not separate social networking with administrators from social networking with everyone either, whereas an administrator (or a bureaucrat) may be more influent than an unknown user on an RfA result, as showed by (Lee et al., 2012).

4 Data collection and model

To measure the respective importance of the knowledge contribution and the socialization for being promoted, we separated the profile of a candidate in two parts. A revision part, which focuses on the revision activities of the candidates, and a social part which is based on their social activities. The social part is computed in two social graphs, one considering interaction between the candidate and every user, and one considering only the candidate and administrators. We choose to consider two graphs because of the hypothesis that socialization with administrators, the future “peers”, has more influence on the promotion than socialization with anybody.

4.1 Dataset

For this study, for convenience but also to be able to benchmark our results with the previous studies we presented, we used a dataset given by the Stanford Large Network Dataset Collection⁴ on Wikipedia. We focused on the RfA occurred in

¹The co-reverts acting negatively.

²<http://en.wikipedia.org/wiki/Wikipedia:GRFA>

³the best indicators of an Feature Article, at least for the English version (Dalip et al., 2009) are the length and basic quality of the writing, as it is for open source contribution, where Hofmann and Riehle (2009) found that the simple heuristics are superior to the more complex text-analysis-based algorithms to estimate the size and the importance of a commit in open-source projects.

⁴<https://snap.stanford.edu/data/#wikipedia>

the period of time from 2006-01-01 to 2007-10-01, because this period contains an important number of RfA, and because there were sufficient activities before 2006 to construct the social networks based on user talks. In this considered period, we removed the RfA done several times in a month by a same user. Hence, the resulting number of RfA in the dataset we considered was 1,617, with a success rate of 49.2%.

4.1.1 The variables

In this part, we describe the different features we consider for modeling candidates.

The revision part It is based on user’s revision activities.

From these activities, we extracted the number of revisions/editions they made (variable: *Revision*), the number of distinct pages (*Pages*) they edited and the number of distinct categories (*Categories*) they participated in, and finally, the repartition of their revisions (*Revision_{repartition}*) in order to take into account both the volume and the variety of the revisions⁵.

Then, we assumed that the users’s talks on the discussion pages of the articles are related to their revision activities, and we added three attributes about this talking on pages: the number of distinct pages where the user talked (*TalkPages*), the total number of talks on the articles’s discussion pages (*PageTalks*), and the repartition of the user talks on these pages (*PageTalks_{repartition}*).

The social part We focused on the conversations on the users’ pages, to assess the impact of what happens beside the discussion on the edits, and more generally, beside the interactions regarding the production of knowledge.

We created three weighted and oriented graphs, based on the social interaction, a general one where the nodes are the considered candidate and all the users, named *userSN*, and two specific ones:

- a graph where nodes are the considered candidate and all the (already) admins, named *adminSN*;
- a graph where nodes are the considered candidate and all the (already) bureaucrats, named *burSN*.

For each graph, we computed the attributes that described the characteristics of the node ‘candidate’. These attributes are described bellow. As they are the same for each graph, we only gave one name for each type of attribute, and added a suffix which is the name of the related graph.

The first attribute is the degree of the node (*Degree*), without taking into account the orientation of edges. Then, for more details, we considered 1) the out degree of the node (*outDegree*) which represents the number of distinct users/admins/bureaus to whom the candidate posted a message to on their user page, 2) the in degree of the node (*inDegree*) which is the number of distinct users/admins/bureaus that posted a message on the candidate’s page. Then, we

⁵For this, we calculated the Gini coefficient on the number of revisions of the user by pages. This attribute allows to quantify the inequalities in a distribution. If it approaches 1, it means that the user was mainly focused on few pages among the whole set of pages revised. Inversely, if it approaches 0, it means the user has had an equal revision behavior on every page revised.

considered the total number of messages posted and received by the candidate (*TalksNumber*). It is different from the Degree since the weightings are used here. The graph being oriented, we took also into account the total number of messages posted to users/admins/bureaus pages (*outTalksNumber*) and the total number of messages received by the candidate (*inTalksNumber*). Then, we computed multiple centrality measures on the graphs:

- the closeness centrality attribute (*Closeness*). This metric is the inverse of the sum of the distances from this node to all the other nodes. The more central a node is, the lower is its total distance to all other nodes. It can be seen as a measure of how long it will take to spread information from the candidate node to all other nodes sequentially (Beauchamp, 1965);
- the PageRank centrality (*PageRank*) (Page et al., 1999), a classic metric on graphs which gives an indicator on whether the candidate node is centric in the graph.
- the betweenness centrality (*Betweenness*) which is equal to the number of shortest paths from all nodes to all others that pass through the candidate's node. A node with high betweenness centrality has a large influence on the transfer of information through the network, under the assumption that information transfer follows the shortest paths (Freeman, 1977). This latter measure have not been computed on the general graph because of limited computer capacity.

Finally, we computed the Gini coefficient for both the number of messages posted by the candidates (*outTalks_{repartition}*) and the number of messages received by them (*inTalks_{repartition}*). These attributes allow to quantify the repartition of the messages from or for the candidates. As said before, a low value (0) means a dispersion behavior whereas a high value means a focused one.

4.2 The models

We created multiple predictive models of RfA success based on the random forest algorithm. This algorithm is a learning method for classification (and regression) that operates by constructing a multitude of decision trees (Quinlan (1993) during training time, and outputting the class that is the dominant value (mode) of the classes output by individual trees. More details can be seen in Breiman (2001).

Each predictive model considered a different modeling of candidate profiles, taking into account subsets of features from the modeling proposed in the previous section. Since we want to understand the contribution of the social attributes in the RfA result, we first created two predictive models, one based on the revision attributes and one based on the social attributes. Then, we considered a model using every attributes. The different types of profiles for each model are described below:

1. The profile based on the revision variables used:
Revisions, Pages, Categories, TalkPages, PageTalks, Revision_{repartition}, PageTalks_{repartition}

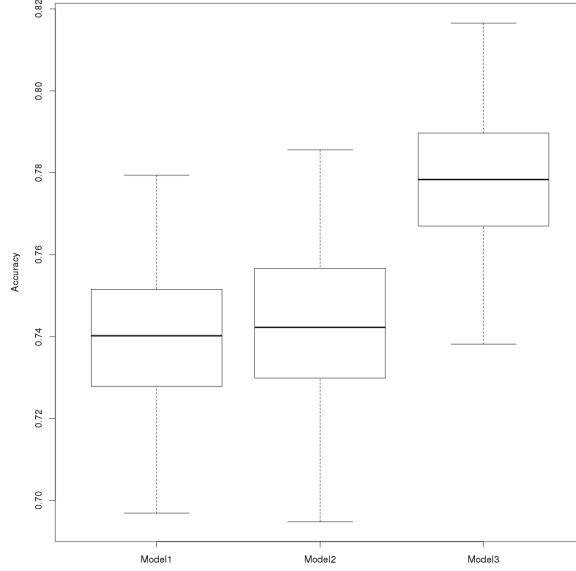


Figure 1: Prediction accuracy for each predictive model

2. The profile based on the social variables used:

$Degree_{adminSN}$, $outDegree_{adminSN}$, $inDegree_{adminSN}$, $TalksNumber_{adminSN}$, $outTalksNumber_{adminSN}$, $inTalksNumber_{adminSN}$, $Closeness_{adminSN}$, $PageRank_{adminSN}$, $Betweenness_{adminSN}$, $outTalksRepartition_{adminSN}$, $inTalksRepartition_{adminSN}$, $Degree_{userSN}$, $outDegree_{userSN}$, $inDegree_{userSN}$, $TalksNumber_{userSN}$, $outTalksNumber_{userSN}$, $inTalksNumber_{userSN}$, $Closeness_{userSN}$, $PageRank_{userSN}$, $outTalksRepartition_{userSN}$, $inTalksRepartition_{userSN}$, $Degree_{burSN}$, $outDegree_{burSN}$, $inDegree_{burSN}$, $TalksNumber_{burSN}$, $outTalksNumber_{burSN}$, $inTalksNumber_{burSN}$, $Closeness_{burSN}$, $PageRank_{burSN}$, $Betweenness_{burSN}$, $outTalksRepartition_{burSN}$, $inTalksRepartition_{burSN}$

3. The profile based on both the social and the revision parts, which used the whole set of variables of the two firsts.

For each predictive model, we separated the dataset in a train and a test sets. The train set consisted in a random 70% of all the candidates and the test set contained the remaining 30%. Then, the predictive model was trained on the train set and applied on the test set to predict RfA success. We compared those predictions to the real RfA success value of the test set, to deduce an accuracy value for each predictive model. Accuracy is the ratio of the number of good predictions on the number of predictions. This process was done 100 times to smooth extreme cases. We present the boxplots of the results in Figure 1.

There are many attributes in model 3 and some of them may be useless, being very correlated to others. Hence, we calculated the Pearson correlation on all attribute's pairs and we removed one element of the pair for which the absolute value of the correlation is over 0.8. More precisely, the attributes: Pages, PageTalks, Degree_*SN, outTalksNumber_userSN, inTalksNumber_userSN,

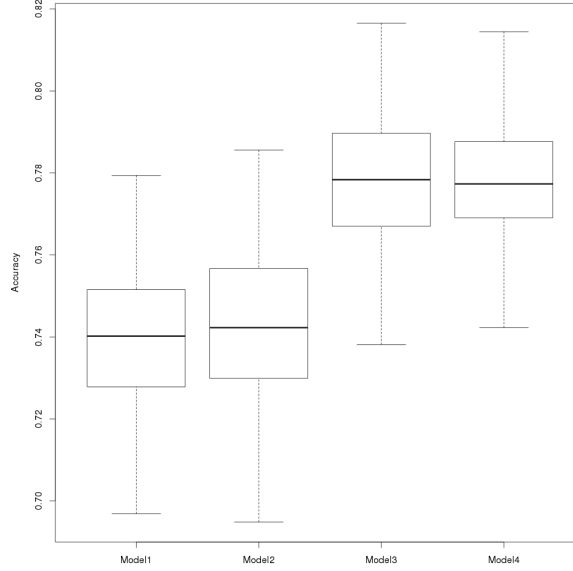


Figure 2: Prediction accuracy with model 4

Table 1: Confusion matrix for each predictive model

		Confusion Matrix		
		0	1	Accuracy
Model 1	0	169	77.5	68.6%
(Revisions)	1	48	190	79.8%
Model 2	0	168	80	67.7%
(Social)	1	46.5	192	80.5%
Model 4	0	177	70	70.6%
(Social + Revisions)	1	37	200	84.4%

TalksNumber_*SN, inTalksNumber_adminSN, outTalksNumber_adminSN, Closeness_*SN, Betweenness_*SN are very correlated to other ones.

After this operation, we obtained a new model (Model 4) which is as good in prediction and better in variance (Figure 2).

Table 1 details the accuracy of the models in predicting either an unsuccessful (0) or a successful (1) promotion, giving the confusion matrix for model 1, model2, and model4.

5 Results

The predictive model based on revision attributes (model 1) and the one based on social attributes (model 2) are almost equivalent in terms of quality of RfA results prediction : median accuracies are respectively 74.0% and 74.2%. In detail, following to the confusion matrices of model 1 and model 2, the accura-

cies are respectively 79.8% vs 80.5% when predicting successful promotions and 68.6% vs 67.7% when predicting unsuccessful ones.

When aggregating social and revisions (Model 4), median prediction accuracy rises up to 77.8%, while the accuracy is 84.4% for predicting successful promotions and 70.6% for predicting unsuccessful ones.

According to these results, social and revisions attributes seem to be complementary for predicting RfA results.

The random forest method gives values for quantifying the importance of an attribute for the prediction quality. For this purpose, it computes the average decrease of accuracy of each tree into the forest when a given attribute is not used. Higher this value is, more important this attribute is for the prediction. Figure 3 presents these attributes by ascending order of importance for Model 4.

In this figure, we can see that the most important attributes in general are Revisions, TalkPages, outDegree_userSN. In particular, to predict the successful promotions, these three attributes are also standing out but not in the same order: outDegree_userSN, Revisions, TalkPages. To predict the unsuccessful promotions, two attributes stand out: Revisions and TalkPages.

These results do not give any information on the preferred values for each attribute. For this purpose, we first compared the density of probabilities of every attribute between accepted candidates and rejected ones. Figure 4 shows the densities of probabilities for the 9 attributes with the most mean decrease accuracy importance.

On this Figure, each plot, but the PageRank plots, highlight significant behavior differences between the promoted and the non-promoted candidates: the interdecile range of the density of probabilities of the promoted candidates is smaller than the non-promoted candidates one (the curve is flatter for non-promoted ones). This suggests that the promoted candidates behave more similarly than the non-ones. This is an explanation of the better prediction of promotion than of no-promotion in Model 4. Moreover, we can see in the first 6 plots that the peak of density for the promoted candidates has bigger value than those for the non-promoted: the successful candidates are more active than the unsuccessful ones.

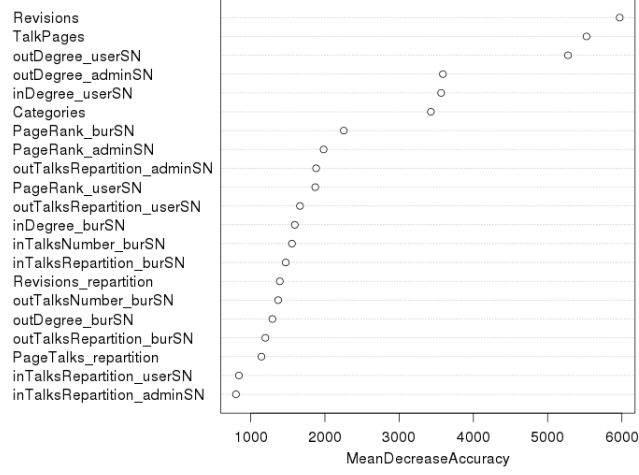
To estimate these differences, we computed the estimated probability of being promoted knowing only each attribute, one by one, for the top 9 attributes (Figure 5 presents the results).

The estimated probabilities of being elected according to the number of revisions (*Revisions*) indicates that the probability of being promoted exceeds 50% beyond about 2,500 revisions and is about 70% beyond 6000 revisions⁶.

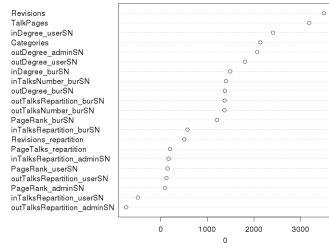
There is a similar behavior pattern (probability of 50% beyond a threshold (T) and extreme values with not enough case studies) for the probabilities of being promoted according to the following attributes: *Categories* ($T \approx 1700$), *TalkPages* ($T \approx 130$), *outDegree_userSN* ($T \approx 450$), *outDegree_adminSN* ($T \approx 17$), *inDegree_userSN* ($T \approx 140$).

The attribute *outTalksRepartition_adminSN* shows that too dispersed candidates in terms of number of talks to admins (thus in the adminSN graph) reduce

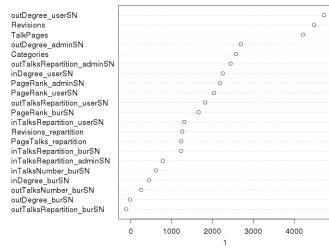
⁶It reaches extreme values (0%, 100%) beyond 40000 revisions because there are not enough case studies of candidates with such big number of revisions. We cannot deduce preferred values on the revision number based on these extreme cases. However, they show that even with a huge number of revisions, a candidate can be rejected.



(a) Importance of attributes for prediction accuracy



(b) Importance of attributes for predicting unsuccessful promotions



(c) Importance of attributes for predicting successful promotions

Figure 3: Importance of attributes in predictive model 4

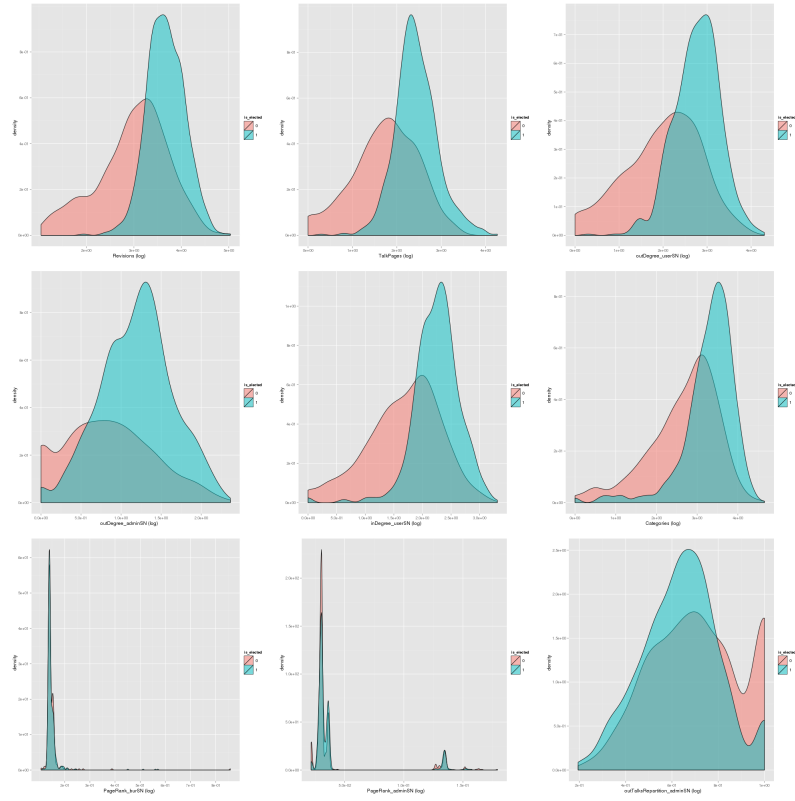
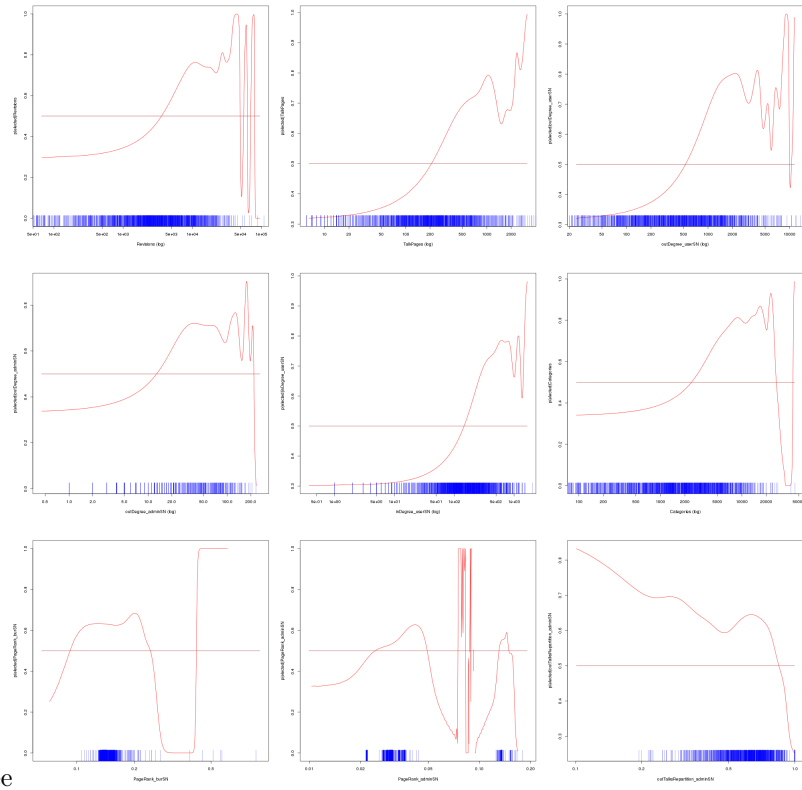


Figure 4: Density of probabilities of each attribute for promoted and non-promoted candidates



three

Figure 5: Estimated probabilities of being promoted according to the top 9 attributes

their chance to be promoted.

No explicit trend is emerging from the probability according to the PageRank attributes. We cannot advise a preferred user social networking behavior with administrator or bureaucrats based on this information

6 Discussion

Our results are consistent with the Guide to RfA, with previous results and with the theories on epistemic communities. Regarding the guide, we provide much more precise figures of how many contributions and interactions are needed to have a good probability to be elected, and we show that there are quite narrow windows in terms of number of contributions and discussions which maximize the chances. Simplifying the measures proposed by [Burke and Kraut \(2008\)](#) regarding the edit activities, and adding the activity on social networks leads to a better evaluation of the chances to be elected (75,6% of good prediction for their model, 78% for our model), keeping the number of explanatory variables reasonably low.

This said, and as pointed out by [Zhu et al.](#)'s study (2011) on the differences between administrative persons ("admin" or "sysop") and project leaders, in the English Wikipedia, who showed that if local project leaders leave more task oriented messages when administrators are more in the social exchange, sending more personal messages in users' personal pages (p. 4): the social activity in the whole project (Talk and userSN variables) is also very important to become administrator. We also refined [Antin et al. \(2012\)](#)'s findings that people involved from the beginning in more diverse revision activities are more likely to take administrative responsibilities. To be elected, a candidate has to be involved in various talkpages, in various articles, but not too much! Being too dispersed (not focused) in terms of revisions or being excessive in the number of contacts (too many) lead to a failure.

Regarding the general discussion about how communities of creation work and can be connected with the study of [Rullani and Haefliger \(2013\)](#) on how core developers are recruited in open source software communities. As supposed for an epistemic community, the contribution in knowledge (Revision) is the first criterion to be considered as a good candidate, and most of the non-elected are not because of a lack of contribution. But, once the candidates have proven their competency (production of knowledge) and their willing to do the job (interacting with people), and even when the choice is opened to no-core members as it is in Wikipedia, knowing and be known by these core members, the future peers, makes the difference. In our case, the variable $outDegree_{adminSN}$ comes fifth in the most important explanatory variables for a positive election, and $inDegree_{burSN}$ seventh.

There are obvious limitations to our work. First, our survey addresses only one project (English Wikipédia) and should be extended to other languages and other epistemic communities. However, as we already pointed out, they are consistent with those found in open source software communities. Second, if our model is good at forecasting the election (more than 80%), it is less good for the non-election (around 70%). Dropping the extreme cases, the people who are not elected because they talked too much, maybe because they fought with the administrators, for instance, may improve the prediction.

But as it is, it already has very practical consequence for the managers and the people involved in those communities. Promoting and encouraging people to take responsibilities in voluntary organization is a major issue for the convenors of communities, being online or offline. Our results suggest that much precise criterion could be posted on what is expected from the candidates to administrative functions in those knowledge production-oriented communities (epistemic communities).

7 Conclusion

In this article, for predicting the success in being promoted in communities of creation, we proposed to consider the social behavior of those candidates in addition to the usual knowledge behavior, looking at their interaction with the participants and with their future peers. We did so looking at the election of administrators in the English Wikipedia.

We compared three predictive models using the random forest algorithm, a revision-based model, a social-based model and mixed model based on social and revisions. We show that using only on social behavior of candidates, it is possible to predict the promotion results with a 74.2% accuracy whereas it is 74.0% considering only their revision behavior. Combining social and revision behaviors, we obtain a 77.8% accuracy which is a little better than the 75.6% accuracy given in [Burke and Kraut \(2008\)](#).

Beyond the predictive model, this article provided estimated probabilities of being elected according to each attribute. They highlighted thresholds under which the candidates reduce their chance to be promoted, but also show that too active candidates in terms of contribution or social interaction may also find it difficult to be elected. Finally, another interesting result is that candidates being too dispersed (not focused) in terms of revisions (many pages with few revisions) and also in terms of social talks (few talks with many users/admins) reduce their chance to be promoted.

In other words, and even if our results must be confirmed in other online epistemic communities, the candidates for responsibilities in those communities must be aware that beyond the “professional” skills requested to be considered for such promotion, taking responsibilities in those big communities mean working in a team, and that the social skills, the knowledge of the incumbents, matter too, as social interaction and coordination are key for the team to be effective, and thus efficient.

References

- J. ANTIN, C. CHESHIRE, and O. NOV**, 2012. Technology-mediated contributions: Editing behaviors among new wikipedians. In *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work, CSCW '12*, pages 373–382, New York, NY, USA. ACM.
- K. BALK, C. RAASCH, and C. HERSTATT**, 2009. Open source enters the world of atoms: A statistical analysis of open design. *First Monday*, 14(11).

- M. A. BEAUCHAMP**, 1965. An improved index of centrality. *Behavioral Science*, 10(2):161–163.
- L. BREIMAN**, 2001. Random forests. *Machine learning*, 45(1):5–32.
- M. BURKE** and **R. KRAUT**, 2008. Mopping up: Modeling wikipedia promotion decisions. In *Proceedings of the 2008 ACM conference on Computer supported cooperative work*, pages 27–36. ACM, 2008.
- B. BUTLER**, **E. JOYCE**, and **J. PIKE**, 2008. Don’t look now, but we’ve created a bureaucracy: The nature and roles of policies and rules in Wikipedia. In *Proceeding of the twenty-sixth annual SIGCHI conference on Human factors in computing systems*, CHI ’08, pages 1101–1110, New York, NY, USA. ACM.
- D. CARDON**, 2012. Discipline but not punish. The governance of Wikipedia. In **F. MASSIT-FOLLÉAT**, **C. MÉADEL**, and **L. MONNOYER-SMITH**, editors, *Normative Experience in Internet Politics*. Presses des Mines, Paris.
- P. COHENDET**, **F. CRÉPLET**, and **O. DUPOUET**, 2001. Interactions between epistemic communities and communities of practice as a mechanism of creation and diffusion of knowledge. In **J.-B. ZIMMERMANN** and **A. KIRMAN**, editors, *Interaction and Market Structure*. Springer, Londres.
- D. H. DALIP**, **M. A. GONÇALVES**, **M. CRISTO**, and **P. CALADO**, 2009. Automatic quality assessment of content created collaboratively by web communities: A case study of wikipedia. In *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries*, pages 295–304, 2009.
- K. EDWARDS**, 2001. Epistemic Communities, Situated Learning and Open Source Software Development. In *Cultures and the Practice of Interdisciplinarity’ Workshop at NTNU*, page 24, 2001.
- Y. FANG** and **D. NEUFELD**, April 2009. Understanding Sustained Participation in Open Source Software Projects. *Journal on Management Information Systems*, 25(4):9–50.
- L. FLEMING** and **D. M. WAGUESPACK**, 2007. Brokerage, Boundary Spanning, and Leadership in Open Innovation Communities. *Organization Science*, 18(2):165–180.
- D. FORAY** and **J.-B. ZIMMERMANN**, October 2001. L’économie du logiciel libre: organisation coopérative et incitation à l’innovation. *Revue économique*, 52:77–93. Special Issue, hors série: économie d’Internet, sous la direction d’É. Brousseau et N. Curien.
- A. FORTE** and **A. BRUCKMAN**, 2005. Why do people write for Wikipedia? Incentives to contribute to open-content publishing. *working paper*.
- L. C. FREEMAN**, 1977. A set of measures of centrality based on betweenness. *Sociometry*, 40:35–41.
- H. HASAN** and **C. PFAFF**, 2006. The wiki: an environment to revolutionise employees’ interaction with corporate knowledge. In *Proceedings of the Australasian Computer-Human Interaction Conference*, Sydney. OZCHI 2006.

- C. HESS** and **E. OSTROM**, 2006. Introduction: An Overview of the Knowledge Commons. In (?), editor, *Understanding Knowledge as a Commons. From Theory to Practice*, pages 3–26.
- P. HOFMANN** and **D. RIEHLE**, 2009. Estimating Commit Sizes Efficiently. In *Open Source Ecosystems: Diverse Communities Interacting (IFIP 2.13)*, volume 299/2009 of *IFIP Advances in Information and Communication Technology*, pages 105 – 115. Springer, Springer, 2009. URL: <http://flosshub.org/sites/flosshub.org/files/EstimatingCommitSizesEfficiently.pdf>.
- T. IBA**, **K. NEMOTO**, **B. PETERS**, and **P. A. GLOOR**, 2010. Analyzing the creative editing behavior of wikipedia editors through dynamic social network analysis. In *Procedia - Social and Behavioral Sciences*, volume 2, pages 6441–6456, 2010.
- M. JANKOWSKI-LOREK**, **L. OSTROWSKI**, **P. TUREK**, and **A. WIERZBICKI**, 2013. Modeling wikipedia admin elections using multidimensional behavioral social networks. *Social Network Analysis and Mining*, 3(4):787–801.
- C. JENSEN** and **W. SCACCHI**, 2007. Role Migration and Advancement Processes in OSSD Projects: A Comparative Case Study. In *ICSE 07: Proceedings of the 29th international conference on Software Engineering*, pages 364–374, Washington, DC, USA. IEEE Computer Society.
- N. JULLIEN**, **K. ROUDAUT**, and **S. LE SQUIN**, novembre-décembre 2011. L’engagement dans des collectifs de production de connaissance en ligne. Le cas GeoRezo. *Revue française de socio-économie*, 8(2):59–83.
- B. LATOUR** and **S. WOOLGAR**, 1979. *Laboratory Life: The Social Construction of Scientific Facts*. Sage Publications, Beverly Hills.
- J. B. LEE**, **G. CABUNDUCAN**, **F. G. CABARLE**, **R. CASTILLO**, and **J. A. MALINAO**, 2012. Uncovering the social dynamics of online elections. *Journal of Universal Computer Science*, 18(4):487–505.
- J. LERNER** and **J. TIROLE**, June 2002. Some simple economics of open source. *Journal of Industrial Economics*, 50:197–234.
- J. LESKOVEC**, **D. HUTTENLOCHER**, and **J. KLEINBERG**, 2010. Predicting positive and negative links in online social networks. In *Proceedings of the 19th international conference on World wide web*, pages 641–650. ACM, 2010.
- D. MAHR** and **A. LIEVENS**, 2012. Virtual lead user communities: Drivers of knowledge creation for innovation. *Research Policy*, 41(1):167 – 177.
- G. MARWELL** and **P. OLIVER**, 1993. *The Critical Mass in Collective Action: A Micro-Social Theory*. Cambridge University Press, Cambridge.
- A. P. MCAFEE**, 2006. Enterprise 2.0: The Dawn of Emergent Collaboration. *Management of Technology and Innovation*, 47(3).
- O. NOV**, November 2007. What motivates wikipedians? *Communications of the ACM*, 50:60–64.

- P. OLIVER, G. MARWELL, and R. TEIXEIRA**, November 1985. A theory of critical mass interdependence, group heterogeneity, and the production of collective action. *American Journal of Sociology*, 91(3):522–556.
- S. O’MAHONY and F. FERRARO**, 2007. The emergence of governance in an open source community. *Academy of Management Journal*, 50:1059–1106.
- E. OSTROM**, 1990. *Governing the Commons: The Evolution of Institutions for Collective Action*. Cambridge University Press.
- L. PAGE, S. BRIN, R. MOTWANI, and T. WINOGRAD**, 1999. The pagerank citation ranking: Bringing order to the web.
- C. PENTZOLD**, August 2011. Imagining the Wikipedia community: What do Wikipedia authors mean when they write about their "community"? *New Media & Society*, 13(5):704–721.
- J. R. QUINLAN**, 1993. *C4. 5: programs for machine learning*, volume 1. Morgan kaufmann.
- F. RULLANI and S. HAEFLIGER**, 2013. The periphery on stage: The intra-organizational dynamics in online communities of creation. *Research Policy*, 42(4):941–953.
- W. SCACCHI**, 2007. Free/open source software development: recent research results and emerging opportunities. In *ESEC-FSE companion 07: The 6th Joint Meeting on European software engineering conference and the ACM SIGSOFT symposium on the foundations of software engineering*, pages 459–468, New York, NY, USA. ACM.
- S. K. SHAH**, July 2006. Motivation, Governance, and the Viability of Hybrid Forms in Open Source Software Development. *Management Science*, 52(2): 1000–1014.
- E. VON HIPPEL**, July 1986. Lead users: a source of novel product concepts. *Management Science*, 32(7):791–805.
- G. VON KROGH, S. HAEFLIGER, S. SPAETH, and M. W. WALLIN**, 2012. Carrots and Rainbows: Motivation and Social Practice in Open Source Software Development. *MIS Quarterly*, 36(2):649–676.
- G. VON KROGH, S. SPAETH, and K. R. LAKHANI**, 2003. Community, joining, and specialization in open source software innovation: A case study. *Research Policy*, 32(7):1217–1241.
- M. M. WASKO and S. FARAJ**, 2005. Why Should I Share? Examining Social Capital and Knowledge Contribution in Electronic Networks of Practice. *MIS Quarterly*, 29(1):35–57.
- H. T. WELSER, D. COSLEY, G. KOSSINETS, A. LIN, F. DOKSHIN, G. GAY, and M. SMITH**, 2011. Finding social roles in Wikipedia. In *Proceedings of the 2011 iConference*, iConference ’11, pages 122–129, New York, NY, USA. ACM.

- B. WORTHEN**, 2008. Why most online communities fail. *The Wall Street Journal*, July 16th.
- H.-L. YANG** and **C.-Y. LAI**, 2010. Motivations of Wikipedia content contributors. *Computers in Human Behavior*, 26(6):1377 – 1383.
- X. ZHANG** and **F. ZHU**, 2011. Group size and incentives to contribute: A natural experiment at chinese wikipedia. *The American Economic Review*, 101(4):1601–1615.
- H. ZHU**, **R. E. KRAUT**, **Y.-C. WANG**, and **A. KITTUR**, 2011. Identifying shared leadership in Wikipedia. In *Proceedings of the 2011 annual conference on Human factors in computing systems*, CHI '11, pages 3431–3434, New York, NY, USA. ACM.